Received: 5 July 2020 | Revised: 16 December 2020 | Accepted: 25 January 2021

DOI: 10.1002/pst.2101

MAIN PAPER

WILEY

A robust permutation test for the concordance correlation coefficient

Alan D. Hutson | **Han Yu**

Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA

Correspondence
Han Yu, Assistant Professor of Oncology, Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA.
Email: han.yu@roswellpark.org

Funding information
National Cancer Institute, Grant/Award Number: P30CA016056; U24CA232979-01: U10CA180822

Abstract

In this work, we developed a robust permutation test for the concordance correlation coefficient (ρ_c) for testing the general hypothesis $H_0 : \rho_c = \rho_{c(0)}$. The proposed test is based on an appropriately studentized statistic. Theoretically, the test is proven to be asymptotically valid in the general setting when two paired variables are uncorrelated but dependent. This desired property was demonstrated across a range of distributional assumptions and sample sizes in simulation studies, where the test exhibits robust type I error control in all settings tested, even when the sample size is small. We demonstrated the application of this test in two real world examples across cardiac output measurements and endocardiographic imaging.

KEYWORDS

measures of agreement, non-normal, small sample, studentization

1 | INTRODUCTION

Measurement of agreement is an essential task in biology and medicine. The often encountered question is whether measurements by two different methods on the same samples produce essentially the same results. For example, it may be of interest to evaluate whether a new assay can reproduce the results of a traditional gold-standard assay for measuring tumor biomarkers in serum, or whether two pathologists have the same ratings on a set of samples for a cancer diagnosis. The measurement of agreement consists of two aspects: accuracy and precision. Accuracy pertains to whether the observed value agrees with the true value systematically, while precision measures the extent to which the observed values conform.[1]

Specific agreement measurements have been designed for different types of data. For categorical data with two levels, McNemar's test is typically used to assess the systematic difference between two measurements. A significant test result would suggest two measurements deviate in a systematic manner. Cohen's κ is a single value measurement for agreement between categorical variables, which is defined as the difference between observed and expected agreement by chance.[2] The approach can be extended to ordinal data with more than two categories by using appropriate weighting schemes.[3] For continuous data, the paired-sample t-test can be used to measure the systematic differences between paired observations. The Bland and Altman diagram plots the difference between two measurements against their means, so as to visualize the pattern and extent of agreement relative to the overall variation.[4] The intraclass correlation coefficient (ICC) can be used as single value measurement of agreement, which represents the between-pair variance as a proportion of the total variance of the observations.[5]

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. Pharmaceutical Statistics published by John Wiley & Sons Ltd

Lin (1989)[6] proposed the concordance correlation coefficient (CCC), which is a widely used and highly cited agreement index between pairs of continuous measurements. Several R packages are available for its calculation, such as DescTools, agRee and cccrm.[7-9] The CCC evaluates the agreement between two readings by measuring the variation of their linear relationship from the 45° line through the origin (the concordance line). It can be expressed as the product of the Pearson correlation coefficient ρ, which measures precision, and a measurement of accuracy, which is a function of the means and standard deviations. This is an advantage of the CCC over other measurements since it evaluates precision and accuracy simultaneously in a single measure. The CCC has also been extended to be modeled as a function of covariates[10,11] and a measure of overall agreement among multiple raters.[12] Nonparametric tests have also been developed to assess the multi-rater agreements based on the CCC.[13]

The CCC has become a popular tool for measuring agreement. Hypothesis testing on the CCC ($H_0 : \text{CCC} = \text{CCC}_0$) is important in assessing whether there is sufficient agreement between two measurements. The test is typically based on the asymptotic distribution of either $\hat{\rho}_c$ or the Z-transformed statistic.[6] It has been widely used in real world applications,[14-17] and has been implemented in the Stata CONCORD module.[18,19] Both asymptotic tests rely on large sample sizes and typically fail to control the Type I error at the desired level when n is small. Under such scenarios, permutation tests provide a strong alternative testing approach. To our knowledge, only limited work has been reported on permutation test about the CCC. Williamson et al.[20] proposed a permutation test for the CCC for comparing whether two methods have equal agreement with the third, for example a gold standard. However, permutation tests for a point null were not a part of their work.

A common and naive mistake in terms of permutation testing about the correlation coefficient or the CCC is to perform a simple permutation test ignoring possible dependency structures, which leads to invalid inference. Since the CCC can be decomposed into the product of Pearson's correlation with a quantity measuring bias, inference about these two measurements are closely related. DiCiccio and Romano have shown that the permutation distribution of Pearson's correlation coefficient does not converge to the sampling distribution when two random variables are dependent but uncorrelated.[21] Therefore, the type I error rate will not be controlled at the desired level. They showed that this issue can be solved by using a permutation test based on an appropriately studentized statistic.

In Section 2 we show that a naive permutation test about the CCC behaves similarly to the non-studentized permutation test about Pearson's correlation coefficient in terms of inflated Type I error rates. To address this issue we propose a studentized statistic for the CCC following the approach of DiCiccio and Romano.[21] More importantly, we extended the studentized permutation test to more general null hypotheses: $H_0 : \text{CCC} = \text{CCC}_0$. Studentized statistics have been widely used in permutation tests.[22,23] However, to our knowledge, this is the first work using studentized permutation test for the CCC. We prove theoretically that the permutation test for the CCC based on studentized statistic is asymptotically valid. In Section 3 we carry out an extensive simulation study which illustrated that studentized permutation test controls the Type I error at its nominal level even in the small sample size settings. Finally, in Section 4 we demonstrate our methodology using real world data from studies on cardiac output measurements and endocardiographic imaging.

2 | METHODS

2.1 | Concordance correlation coefficient

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n pairs of samples independently selected from a bivariate population with means μ_1 and μ_2, and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

The CCC is based on the expected value of the squared difference of two variables, X and Y, and defined as,

$$\rho_c = 1 - \frac{E\left[(X-Y)^2\right]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2},$$

where $-1 \leq \rho_c \leq 1$. Note that the CCC can be decomposed into two parts as follows:

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \rho C,$$

where ρ is the Pearson correlation coefficient, which measures linear association between X and Y, and

$$C = \frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2},$$

which is a measure of accuracy, and represents how far the best-fit line deviates from the 45° line through origin (concordance line). The value of $\rho_c = 1$ indicates perfect agreement while the value of $\rho_c = 0$ indicates lack of agreement. It is important to note that $\rho_c = 0$ if and only if $\rho = 0$.

The null hypothesis that is of interest to most researchers in the context of testing agreement is $H_0 : \rho_c = \rho_{c(0)}$ and we will focus on one-sided alternative hypothesis $H_1 : \rho_c > \rho_{c(0)}$. In agreement testing the test $H_0 : \rho_c \geq \rho_{c(0)}$ versus $H_1 : \rho_c < \rho_{c(0)}$ is generally not of interest in practice. For n independent sample pairs (X_1, Y_1), ..., (X_n, Y_n), ρ_c can be estimated by replacing the population quantities with the respective moment estimators such that

$$\hat{\rho}_c = \frac{2\hat{\sigma}_{12}}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + (\hat{\mu}_1 - \hat{\mu}_2)} = \hat{\rho}\hat{C}$$
$$\hat{C} = \frac{2\hat{\sigma}_1\hat{\sigma}_2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + (\hat{\mu}_1 - \hat{\mu}_2)^2},$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the sample means, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the sample variances, for X and Y, respectively and $\hat{\sigma}_{12}$ is the sample covariance. Note that $\rho_c = 0$ if and only if $\rho = 0$. However, testing $H_0 : \rho_c = 0$ and $H_0 : \rho = 0$ are different tests in that the estimator $\hat{\rho}$ is scaled by a random variable $\hat{C}$ in $\hat{\rho}_c$ compared with $\hat{\rho}$, for the respective tests.

The test about $\hat{\rho}_c$ can be performed based on the asymptotic normal distribution of either $\hat{\rho}_c$ or using the Fisher's Z transformation given as,

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2}\ln\frac{1 + \hat{\rho}_c}{1 - \hat{\rho}_c}.$$

The variances for each test are obtained utilizing the delta method, for example, see Lin (1989)[6] for details of deriving the asymptotic distributions for both statistics, respectively. When the sample size is small, the Type I error is usually not well controlled, even though the Z-transformed statistic converges at a faster rate to normality.[6] We will illustrate this property in our simulation study in Section 3.

2.2 | Background information

As noted earlier the CCC, ρ_c, may be decomposed into two components, namely ρ rescaled by a non-zero constant C. Therefore, the tests of ρ_c and ρ are closely related. In this section, we start by reviewing the permutation test for Pearson's correlation coefficient as developed by Diciccio and Romano.[21] Towards this end define $\mathbf{G}_n$ to be the set of all permutations π of $\{1, ..., n\}$. For testing independence between two random variables X and Y, the permutation distribution of any given test statistic $T_n(X^n, Y^n)$ is defined as

$$\hat{R}_n^{T_n}(t) = \frac{1}{n!}\sum_{\pi \in G_n} I\left\{T_n\left(X^n, Y_\pi^n\right) \leq t\right\}, \tag{1}$$

where Y_π^n represents $\{Y_{\pi(1)}, ..., Y_{\pi(n)}\}$. In this setting, the permutation $\mathbf{G}_n$ is all possible pairwise combinations between X^n and Y^n. A level α one-sided permutation test rejects if $T_n\left(X^n, Y_\pi^n\right)$ is larger than the $1 - \alpha$ quantile of the permutation distribution. The permutation test is exact when exchangeability assumptions hold, that is, the distribution of (X^n, Y^n) is

invariant under the group of transformations $\mathbf{G}_n$. The test using the Pearson correlation coefficient $\hat{\rho}$ is exact when using a metric of dependence for testing the null hypothesis of independence given as

$$H_0 : P = P_X \times P_Y,$$

where P_X and P_Y are marginal distributions. The null hypothesis of independence is not equivalent to the test about zero correlation given as $H_0 : \rho = 0$ with the exception of limiting assumptions such as the data are distributed as bivariate normal random variables. In other words, in the general setting two random variables can be dependent but uncorrelated. In such cases, DiCiccio and Romano[21] have shown that, with finite fourth moments, the permutation distribution of $\hat{\rho}$ converges to $N(0, 1)$, but its sampling distribution converges to $N(0, \tau^2)$, where

$$\tau^2 = \frac{\mu_{22}}{\mu_{20}\mu_{02}},$$

and

$$\mu_{rs} = E[(X_1 - \mu_1)^r (Y_1 - \mu_2)^s].$$

Thus the test will not be level α unless $\tau = 1$. In light of this result, DiCiccio and Romano proposed a studentized correlation test statistic, which has been shown to control Type I error asymptotically at α when two random variables are dependent but uncorrelated.[21] Specifically, the studentized statistic is defined as $S_n = \sqrt{n}\hat{\rho}_n / \hat{\tau}_n$, where

$$\hat{\tau}_n^2 = \frac{\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}},$$
$$\hat{\mu}_{rs} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^r (Y_i - \bar{Y})^s.$$

The permutation distribution and sampling distribution of S_n both converge to the standard normal distribution asymptotically. It should be noted that even though the results presented in DiCiccio and Romano[21] are based on large sample approximations the behavior of this test for small to moderate sample sizes is quite good as born out in their simulation results.

2.3 | Permutation concordance correlation test for $H_0 : \rho_c = 0$

For the permutation test of $\rho_c = 0$, we use the same permutation scheme used for the Pearson correlation coefficient[21] as described in Equation (1). That is, for each permutation we will randomly shuffle Y while keeping X fixed. Recall that $\rho_c = 0$ if and only if $\rho = 0$, but we also have $\rho_c \to 0$ when $\sigma_1/\sigma_2 \to +\infty$ or 0, or when $|\mu_1 - \mu_2| \to +\infty$. The latter condition implies F_x and F_y have either location or scale differences or both. Therefore, for any permutation scheme, the exchangeability assumption does not necessarily hold under H_0. In this section we show that the permutation test using the permutation scheme by randomly shuffling Y, although not exact, will be asymptotically valid if the statistic is properly studentized. On the other hand, the naive permutation test based on the statistic $\hat{\rho}_c$ defined at Section 2.1 suffers a similar deficiency to that of $\hat{\rho}$ for Pearson's correlation coefficient, and it does not generally control the Type I error at the desired level.

From the result of DiCiccio and Romano,[21] if $E(X_1^2) < \infty, E(Y_1^2) < \infty$ and $E(X_1^2 Y_1^2) < \infty$, then under H_0, $\sqrt{n}\hat{\rho}_n \to \tau Z$, where $Z \sim N(0, 1)$. By the strong law of large numbers, we have $\hat{C}_n \to C$ almost surely. Therefore, by Slutsky's theorem, we have $T_n^c = \sqrt{n}\hat{\rho}_n \hat{C}_n \to \tau C Z$ in distribution under H_0.

Proposition 1. *Let C_n and T_n be functions of a sequence of i.i.d. random variables X_n. If $C_n \to C$ almost surely, and*

$$\lim_{n\to\infty} \sup_{t\in R} |\hat{R}_n^{T_n}(t) - F(t)| = 0$$

for almost every sequence of X_n, where F(t) is the CDF of Z, then for statistic $T_n' = C_n T_n$, we have

$$\lim_{n\to\infty}\sup_{t\in R}|\hat{R}_n^{T_n'}(t)-F'(t)|=0,$$

under $H_0:\rho_c=0$, *where* $F'(t)$ *is the CDF of* $Z'=CZ$.

Note that for a given sample, $\hat{C}\left(X^n,Y_\pi^n\right)$ remains constant for any π, and $\hat{C}_n\to C$ almost surely. From Theorem 2.1 of DiCiccio and Romano[21] it follows that

$$\lim_{n\to\infty}\sup_{t\in R}|\hat{R}_n^{T_n}(t)-F(t)|=0.$$

Therefore, by Proposition 1, we have,

$$\lim_{n\to\infty}\sup_{t\in R}|\hat{R}_n^{T_n^c}(t)-\Phi_C(t)|=0$$

almost surely, where Φ_C is the CDF of $N(0,C^2)$. We can see that the sampling and permutation distributions of T_n converge to the same distribution only when $\tau=1$, thus the permutation test will not guarantee Type I error control at level α under general scenarios. On the other hand, the permutation test will be asymptotically valid when it is based on a studentized statistic. Towards this end we have the following:

Theorem 1. *Let* (X_n, Y_n) *be a sequence of i.i.d. random variables. Suppose* $E\left(X_1^4\right)<\infty$ *and* $E\left(Y_1^4\right)<\infty$, *and define the studentized statistic*

$$S_n^c=\sqrt{n}\hat{\rho}_n\hat{C}_n/\hat{\tau}_n,$$

then we have

$$\lim_{n\to\infty}\sup_{t\in R}|\hat{R}_n^{S_n^c}(t)-\Phi_C(t)|=0,$$

almost surely under $H_0:\rho_c=0$, *where* Φ_C *is the CDF of* $N(0,C^2)$.

The proof of Theorem 1 follows directly from the fact that C_n is constant under permutations, Theorem 2.2[21] and Proposition 1. Therefore, both the sampling distribution and permutation distribution of $\sqrt{n}S_n^c(X^n,Y^n)$ converge to the corresponding quantiles of a $N(0,C^2)$ distribution, which in turn proves the test has asymptotic Type I error control at level α. Note that although $\hat{C}$ remains constant under the proposed permutation scheme, the tests on the Pearson's correlation coefficient and the CCC are distinctly different tests.

2.4 | Permutation concordance correlation test for $H_0:\rho_c=\rho_{c(0)}$

In a more general scenario, we may be interested in testing $H_0:\rho_c=\rho_{c(0)}$ versus $H_0:\rho_c>\rho_{c(0)}$. This corresponds to a non-zero correlation under H_0, which cannot be tested by a conventional permutation test. In order to bring this test into the above framework, we rely on a statistic based on a de-correlated sample.

First, we obtain an estimated correlation under the H_0:

$$\hat{\rho}_0=\rho_{c(0)}/\hat{C},$$

which converges almost surely to ρ_0 when H_0 is true. We can standardize the original observations by

$$U_i' = \frac{X_i - \bar{X}}{S_X}, V_i' = \frac{Y_i - \bar{Y}}{S_Y},$$

such that the new variables will have zero means and unit standard deviations, which will then be de-correlated by

$$(U,V)^T = \mathbf{A}(\hat{\rho}_0)(U',V')^T.$$

In this equation, the matrix $\mathbf{A}(\cdot)$ is defined as

$$\mathbf{A}(x) = \begin{pmatrix} 1 & 0 \\ \frac{-x}{\sqrt{1-x^2}} & \frac{1}{\sqrt{1-x^2}} \end{pmatrix},$$

which satisfies

$$\mathbf{A}(\rho)\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\mathbf{A}(\rho)^T = \mathbf{I}_2.$$

It is straightforward to show that $\hat{\rho}(U,V) \to 0$ under H_0. The test statistic can then be defined based on U and V,

$$S_n^c = \sqrt{n}\hat{\rho}_n(U,V)\hat{C}_n/\hat{\nu}_n,$$

where $\hat{\rho}_n(U,V)$ is the sample Pearson correlation of U and V, and ν_n^2 is the large sample variance of $\hat{\rho}_n(U,V)$. By large sample theory, it can be readily shown that $\sqrt{n}\hat{\rho}_n(U,V) \to N\left(0,\nu_n^2\right)$. Therefore, we have

$$S_n^c \to N\left(0,C^2\right).$$

Obviously, the pair (U, V) is asymptotically uncorrelated under H_0, which means they are also asymptotically exchangeable under normality. Therefore, under the framework by DiCiccio and Romano,[21] it is legitimate to obtain the permutation distribution of S_n^c by randomly shuffling V. Specifically, for each permutation we will calculate the statistic as

$$S_n^c(U,V_\pi) = \sqrt{n}\hat{\rho}_n(U,V_\pi)\hat{C}_n.$$

By Theorem 1, we can see that the permutation distribution of S_n^c will converge to $N(0, C^2)$ when $\rho(U, V) = 0$, a condition will be satisfied asymptotically.

The computation of S_n^c relies on the estimation of ν_n^2, of which the analytical form is very difficult to derive. The commonly used bootstrap method yields poor variance estimation in our case (data not shown). This is likely due to the standardization step which tends to be unstable when there are many duplicates during the resampling, which can be more severe when the sample size is small. On the other hand, the jackknife method provides a robust estimation. Conventionally, the jackknife procedure calculates $S_{n(i)}^c$ for $i = 1, ..., n$, where $S_{n(i)}^c$ is the statistic with (X_i, Y_i) left out, and the variance can be estimated by $\hat{\nu}_n^2 = (n-1)\hat{\mathrm{Var}}\left(S_{n(i)}^c\right)$. It should be noted that this variance is estimated under $\rho_c = \hat{\rho}_c$. However, for hypothesis testing the variance needs to be estimated under $\rho_c = \rho_{c(0)}$. It is obvious that ν_n^2 depends on ρ_c and such discrepancy may lead to a reduced power. To solve this issue, we used a surrogate distribution approach used in the work by Hutson.[24]

The surrogate data is obtained by a de-correlation operation followed by a re-correlation based on the correlation expected under H_0 which is $\hat{\rho}_0$. Specifically, let U_i' and V_i' be the standardized observations. They will be transformed as below

$$\left(X_i', Y_i'\right)^T = \begin{pmatrix} S_X & 0 \\ 0 & S_Y \end{pmatrix} \mathbf{B}(\hat{\rho}_0)\mathbf{A}(\hat{\rho})\left(U_i', V_i'\right)^T + (\bar{X}, \bar{Y})^T,$$

where $\mathbf{B}(\cdot)$ is defined as

$$\mathbf{B}(x) = \begin{pmatrix} 1 & 0 \\ x & \sqrt{1-x^2} \end{pmatrix}.$$

The matrix $\mathbf{B}(\hat{\rho}_0)\mathbf{A}(\hat{\rho})$ transforms $\left(U_i', V_i'\right)$ into a pair of standardized variables with zero means and unit standard deviations. Importantly, U_i' and V_i' have a correlation of $\hat{\rho}_0$. Through the following rescale and shift operations, the final variable pairs $\left(X_i', Y_i'\right)$ will have the same means and variances as (X_i, Y_i), but a correlation of $\hat{\rho}_0$ instead of $\hat{\rho}$. This ensures the sample ρ_c of $\left(X_i', Y_i'\right)$ is $\rho_{c(0)}$. Therefore, we will use (X', Y') for the jackknife procedure for estimating ν_n^2.

Occasionally, the value of $\hat{\rho}_0 = \rho_{c(0)}/\hat{C}$ could be larger than 1. This will happen when $\hat{C}$ is too small but $\rho_{c(0)}$ too large. In our implementation, we will set $\hat{\rho}_0 = 0.99$ in this scenario, which results in a large p-value. Note that a small C implies large difference between X and Y, which already suggests a poor agreement. Therefore, in such cases it is unreasonable to have large $\rho_{c(0)}$.

3 | SIMULATIONS

3.1 | Test for $H_0 : \rho_c = 0$ versus $H_1 : \rho_c > 0$

We examined the Type I error control using distributions commonly found in the literature for examining these types of test statistics across a wide range of settings.[21,24] For our simulation, we focused on testing $H_0 : \rho_c = 0$ versus $H_1 : \rho_c > 0$, with sample sizes $n = 10, 25, 50, 100, 200$. Each simulation utilized 10,000 Monte Carlo replications and the number of permutations used was 1000. We compared the straight large sample approximation (Asymptotic), Fisher's Z-transformation (Fisher's Z), naive permutation test (Perm), and studentized permutation test (Stu Perm). The Type I error control for $\alpha = 0.05$ was examined. The simulation scenarios from DiCiccio and Romano were utilized in our study:

1. Multivariate normal (MVN) with mean zero and identity covariance.
2. Exponential given as $(X, Y) = rS^T u$ where $S = \text{diag}\left(\sqrt{2}, 1\right)$, u is uniformly distributed on the two dimensional unit circle and $r \sim \exp(1)$.
3. Circular given as the uniform distribution on a two dimensional unit circle.
4. $t_{4.1}$ where $X = W + Z$ and $Y = W - Z$, where W and Z are i.i.d. random variables following t-distributions with 4.1 degrees of freedom.
5. Multivariate t-distribution (MVT) with location parameters $(0, 0)^T$, identity covariance and 5 degrees of freedom.

The results show that for all distributions, both the untransformed and Z-transformed asymptotic tests have inflated Type I error rates when n is small (Table 1). Although the error rates converge towards the nominal level of 0.05 as n increases, they only approach 0.05 when $n \geq 100$. For MVN, the error rates are well controlled by naive permutation test, even when n is small. However, with other distributions, the test is either too conservative (circular) or too liberal (exponential, $t_{4.1}$ and MVT), and the error rates do not converge to 0.05 as n increases. For example, under the $t_{4.1}$ distribution, the type I error rate of naive permutation test inflates dramatically to 0.21 when $n = 200$. Meanwhile, for the circular distribution, this test becomes over conservative with a type I error rate of 0.01 when $n = 200$. On the other hand, the studentized test controls type I error rate robustly at 0.05 under all settings and even when the sample size is as small as 10. The above results demonstrated the proposed studentized permutation test is a robust method for testing $H_0 : \rho_c = 0$. Although $C = 1$ in this setting, the test is not equivalent to $H_0 : \rho = 0$, because the test on ρ_c needs to take into account the variability of $\hat{C}$.

In addition to the original settings, we further examined the tests' type I error control when $C \neq 1$. Let μ_2^0 and σ_2^0 be the mean and standard deviation of Y in the original settings. This is achieved by either introducing a shift in Y, that is $\mu_2 = \mu_2^0 + 2$, or having Y rescaled by a factor of 2, $\sigma_2 = 2\sigma_2^0$. The results are shown in supporting information, which suggest that the studentized permutation test is robust to shifts and different scales in underlying distributions.

TABLE 1 Type I errors for tests of $H_0 : \rho_c = 0$ versus $H_1 : \rho_c > 0$, when $\rho_c = 0$

Distribution	*N*	Asymptotic	Fisher's *Z*	Perm	Stu Perm
MVN	10	0.1227	0.1230	0.0489	0.0457
	25	0.0848	0.0827	0.0516	0.0518
	50	0.0677	0.0660	0.0519	0.0504
	100	0.0599	0.0587	0.0514	0.0516
	200	0.0525	0.0521	0.0479	0.0494
Exponential	10	0.1995	0.1956	0.1157	0.0608
	25	0.1410	0.1317	0.1501	0.0554
	50	0.1107	0.1037	0.1617	0.0544
	100	0.0779	0.0736	0.1619	0.0480
	200	0.0702	0.0673	0.1658	0.0517
$t_{4.1}$	10	0.1665	0.1581	0.0902	0.0444
	25	0.1237	0.1123	0.1313	0.0435
	50	0.0999	0.0911	0.1552	0.0434
	100	0.0904	0.0822	0.1851	0.0487
	200	0.0788	0.0728	0.2051	0.0484
Circular	10	0.0853	0.0907	0.0184	0.0560
	25	0.0589	0.0596	0.0114	0.0473
	50	0.0534	0.0541	0.0101	0.0482
	100	0.0510	0.0512	0.0119	0.0480
	200	0.0503	0.0503	0.0117	0.0478
MVT	10	0.1579	0.1569	0.0721	0.0486
	25	0.1169	0.1114	0.0987	0.0466
	50	0.0878	0.0822	0.1076	0.0436
	100	0.0740	0.0692	0.1175	0.0435
	200	0.0703	0.0679	0.1290	0.0500

Abbreviations: MVN, multivariate normal; MVT, multivariate *t*-distribution.

3.2 | Test for $H_0 : \rho_c = \rho_{c(0)}$ versus $H_1 : \rho_c > \rho_{c(0)}$

Next, we evaluated the proposed test's performance on non-zero null hypotheses. The data was generated under the settings of $\rho_c = 0.3$ or 0.7, $\mu_2 - \mu_1 = 0.5$, $\sigma_2/\sigma_1 = 1.5$. The data was first generated in a similar procedure as the previous section, but was then standardized by the population standard deviations and correlated using $\mathbf{B}(\rho_c/C) = \mathbf{B}(\rho)$. The correlated data was then scaled and shifted to have desired scale and location parameters.

The naive permutation test cannot handle a non-zero point null, thus was not examined. Table 2 shows the type I errors for testing $H_0 : \rho_c = 0.3$ and $H_0 : \rho_c = 0.7$. The results are also shown graphically in Figures 1 and 2. The asymptotic test and the Fisher's Z test generally have inflated type I errors, where the Fisher's Z test shows a faster convergence to 0.05. In fact, the asymptotic test fails to achieve satisfactory type I error control even when n is as large as 200 except for the circular scenario. The test based on Fisher's Z statistic also needs $n > 50$ to achieve good type I error control in most cases. For testing $H_0 : \rho_c = 0.3$ under $t_{4.1}$, the Fisher's Z test fails to control type I error under 0.069 even when $n = 200$. On the other hand, the studentized permutation test robustly controls the type I error at desired level only with a few cases of slight inflation for testing $H_0 : \rho_c = 0.3$ when $n = 10$ (Figure 1). This can be observed for exponential, $t_{4.1}$ and MVT distributions. However, even in these cases, its type I error control is still superior to the competitors. For testing $H_0 : \rho_c = 0.7$, the performance was robust for almost all cases (Figure 2).

We also investigated the power for testing $H_0 : \rho_c = 0.7$ and $H_1 : \rho_c > 0.7$ when the true ρ_c is 0.8 (Figure 3). The exact rejection probabilities are provided in Table S3. It should be noted that in most of the cases, especially when $n \leq 50$, the powers were not comparable because Fisher's Z and asymptotic tests tend to have inflated type I errors. Especially, the asymptotic tests have highest power in all cases, which is due to its inflated type I errors in general. To make legitimate

TABLE 2 Type I errors for tests of $H_0 : \rho_c = \rho_{c(0)}$ versus $H_1 : \rho_c > \rho_{c(0)}$, where $\rho_{c(0)} = 0.3$ or 0.7

		$\rho_{c(0)} = 0.3$			$\rho_{c(0)} = 0.7$		
Distribution	***N***	**Asymptotic**	**Fisher's *Z***	**Stu Perm**	**Asymptotic**	**Fisher's *Z***	**Stu Perm**
MVN	10	0.1195	0.1082	0.0647	0.1219	0.0974	0.0495
	25	0.0853	0.0721	0.0550	0.0888	0.0668	0.0489
	50	0.0692	0.0594	0.0514	0.0757	0.0567	0.0476
	100	0.0632	0.0557	0.0535	0.0638	0.0513	0.0472
	200	0.0508	0.0461	0.0443	0.0572	0.0480	0.0460
Exponential	10	0.1839	0.1660	0.0840	0.1528	0.1264	0.0499
	25	0.1345	0.1120	0.0707	0.1243	0.0912	0.0480
	50	0.1053	0.0865	0.0620	0.0961	0.0684	0.0459
	100	0.0897	0.0741	0.0628	0.0858	0.0614	0.0461
	200	0.0747	0.0662	0.0567	0.0755	0.0577	0.0493
$t_{4.1}$	10	0.1634	0.1491	0.0720	0.1385	0.1085	0.0464
	25	0.1230	0.1061	0.0590	0.1075	0.0699	0.0424
	50	0.1065	0.0880	0.0552	0.0901	0.0589	0.0389
	100	0.0977	0.0813	0.0551	0.0819	0.0515	0.0401
	200	0.0818	0.0690	0.0514	0.0680	0.0464	0.0399
Circular	10	0.0672	0.0606	0.0372	0.0798	0.0628	0.0354
	25	0.0577	0.0511	0.0406	0.0676	0.0508	0.0430
	50	0.0529	0.0464	0.0441	0.0590	0.0467	0.0460
	100	0.0493	0.0455	0.0431	0.0547	0.0452	0.0449
	200	0.0516	0.0480	0.0486	0.0547	0.0477	0.0503
MVT	10	0.1498	0.1334	0.0694	0.1357	0.1097	0.0473
	25	0.1149	0.0968	0.0604	0.1067	0.0794	0.0445
	50	0.0941	0.0773	0.0579	0.0884	0.0629	0.0424
	100	0.0806	0.0684	0.0568	0.0786	0.0587	0.0452
	200	0.0686	0.0595	0.0512	0.0676	0.0527	0.0436

Note: The true ρ_c values are 0.3 and 0.7, respectively.
Abbreviations: MVN, multivariate normal; MVT, multivariate *t*-distribution.

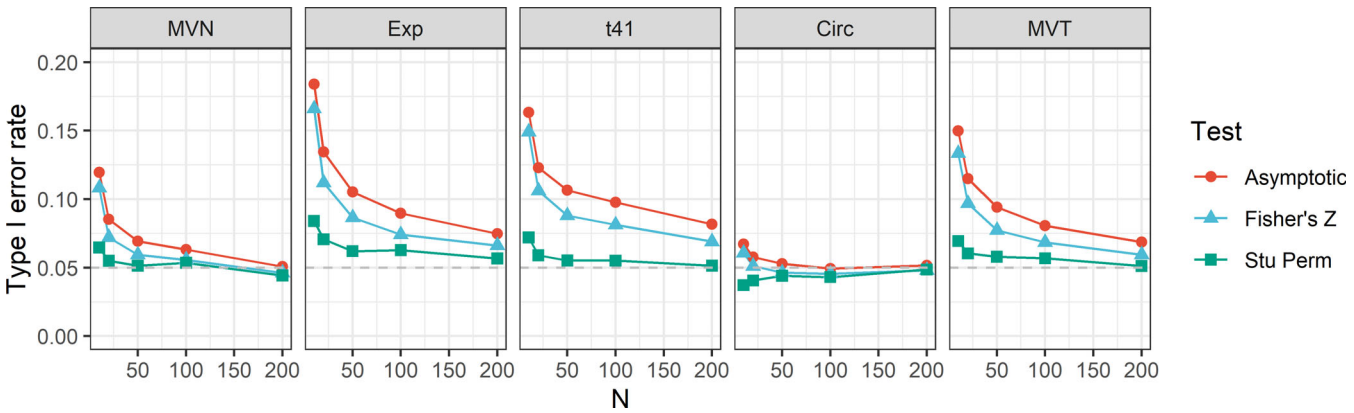

FIGURE 1 Rejection probabilities (type I error) for tests of $H_0 : \rho_c = 0.3$ versus $H_1 : \rho_c > 0.3$, when $\rho_c = 0.3$

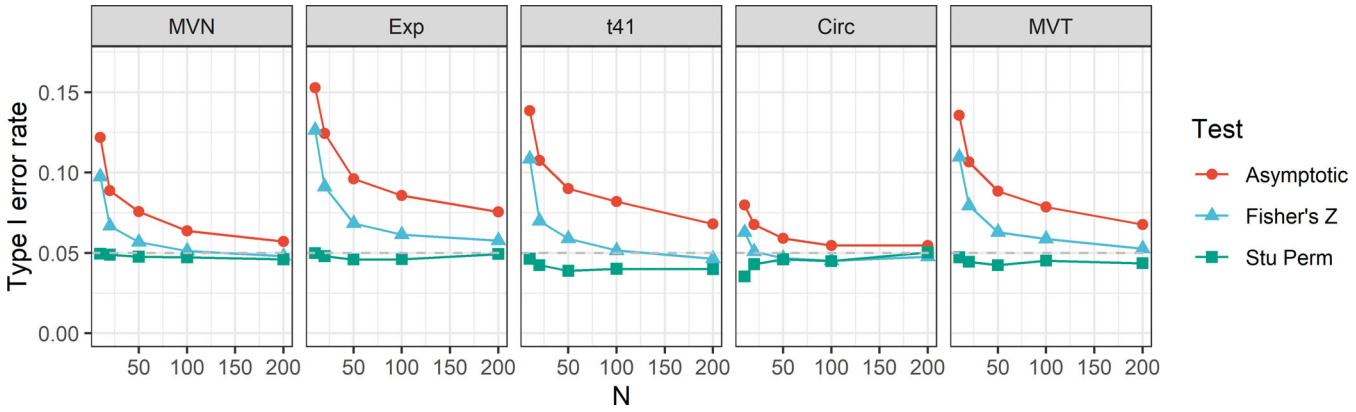

FIGURE 2 Rejection probabilities (type I error) for tests of $H_0 : \rho_c = 0.7$ versus $H_1 : \rho_c > 0.7$, when $\rho_c = 0.7$

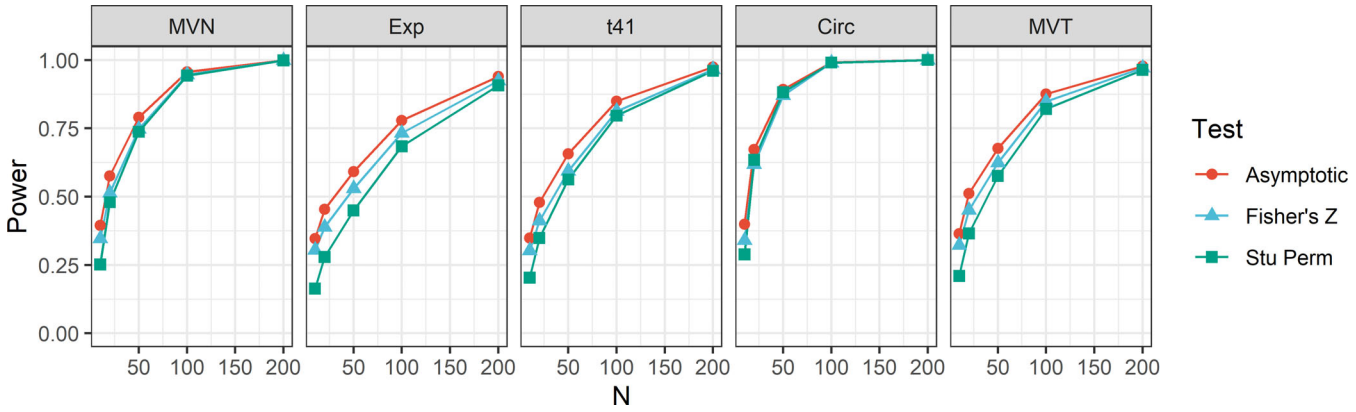

FIGURE 3 Rejection probabilities (power) for tests of $H_0 : \rho_c = 0.7$ versus $H_1 : \rho_c > 0.7$, when $\rho_c = 0.8$

comparisons, we focus on the settings where the type I error of Fisher's Z test is <0.06. In these cases, the difference between the power of Fisher's Z and studentized permutation tests is generally less than 3%. For example, under the circular scenarios with $n \geq 25$, the two tests only show negligible difference in power. Therefore, we can conclude that the proposed test robustly controls the type I error while maintains comparable power with Fisher's Z tests.

4 | EXAMPLE

4.1 | Cardiac output data

As an illustration of our approach, we tested $H_0 : \rho_c = \rho_{c(0)}$ versus $H_1 : \rho_c > \rho_{c(0)}$ using cardiac output estimated from systolic time intervals based on impedance cardiography (IC) and those estimated by radionuclide ventriculography (RV) in 12 patients. The data was originally reported by Bowling et al.[25] and is obtained from the publication by Bland and Altman.[26] In the reported data, there are multiple pairs of observations per patient. Since modeling replicates within individuals is not within the scope of our study, we took average of the replicates for each individual, such that only one pair of averaged measurements for each individual was used for analysis. The scatter plot of the paired data is given in Figure 4. Marginal normality of data was examined by Shapiro–Wilk test, while the bivariate normality was examined by Henze–Zikler test. The p values of Shapiro–Wilk tests for IC and RV are 0.96 and 0.97, respectively. The p value of Henze–Zikler test is 0.99. Therefore, there is no evidence that the data distribution is non-normal.

The estimated $\hat{\rho}_c$ between IC and RV is 0.64. The paired-sample t-test shows cardiac outputs estimated by RV are significantly higher than the estimates by IC ($p = 0.026$), suggesting there is a systematic difference between two

techniques. The p values for $\rho_{c(0)} = 0$ and $\rho_{c(0)} = 0.3$ are shown in Table 3. Both permutation tests used 5000 resamples. If we are testing at level $\alpha = 0.05$, then we will reject $H_0 : \rho_c = 0$ for all tests, and conclude there is a non-zero agreement between IC and RV in estimating cardiac outputs (Table 3). Meanwhile, we are also interested in whether there is a moderate agreement. Therefore, we tested $H_0 : \rho_c = 0.3$ versus $H_1 : \rho_c > 0.3$. In this case, the naive permutation test cannot be applied. Both asymptotic and Fisher's Z tests rejected the H_0. This is a different conclusion from the studentized permutation test, which failed to reject H_0. Based on the simulation results, the studentized permutation test is more reliable when the sample size is small.

4.2 | Echocardiographic imaging

The second example is taken from an echocardiographic imaging (EI) study.[27] The study developed an autonomous boundary detection (ABD) algorithm to detect the limiting boundaries of the left ventricular myocardium, which requires no observer input. The study aimed to increase reliability, objectivity, and reproducibility in order to enhance the quantitative accuracy of echocardiography.

Following the approach by Hutson,[13] we have selected a subset of $n = 15$ subjects from the EI study, and use the fractional area change (FAC) as the quantity of interest. The FAC is computed as

$$\text{FAC} = \frac{A_{\text{ED}} - A_{\text{ES}}}{A_{\text{ED}}} \times 100\%,$$

where A_{ED} and A_{ES} are the endocardial areas at end diastole and end systole respectively. In this example we focus on the comparison between the FAC's as measured by the fuzzy gold standard (FGS) derived from a consensus of experts and one of the echocardiographers (Expert 2), so as to examine the tests when the agreement is poor. The agreement between

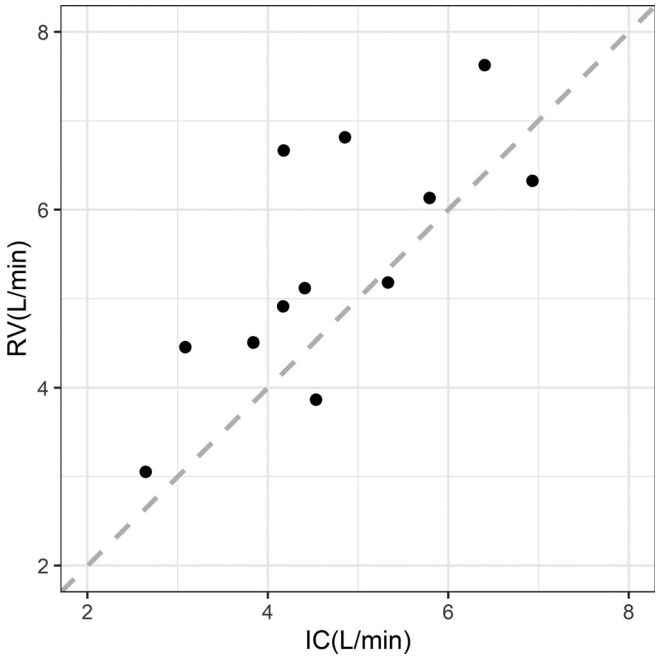

FIGURE 4 Scatter plot of cardiac output data with concordance line

TABLE 3 The p-values of testing $H_0 : \rho_c = \rho_{c(0)}$ versus $H_1 : \rho_c > \rho_{c(0)}$ for cardiac output data

Tests	$\rho_{c(0)} = 0$	$\rho_{c(0)} = 0.3$
Asymptotic	<0.0001	0.0141
Fisher's Z	0.0009	0.0320
Perm	0.0038	–
Stu Perm	0.0146	0.1178

FGS and the expert is visualized by a scatter plot with concordance line (Figure 5). Similarly, marginal normality of data was examined by Shapiro–Wilk test, and the bivariate normality was examined by Henze–Zikler test. The p values of Shapiro–Wilk tests for FGS and expert are 0.20 and 0.09, respectively. The p value of Henze–Zikler test is 0.09.

The paired-sample t-test shows cardiac outputs estimated by FGS are significantly higher than the estimates by the expert ($p = 0.02$) suggesting a systematic difference. From the scatter plot (Figure 5), a low agreement between two measurements was observed, and the CCC is estimated as 0.42. Table 4 shows the results for testing $H_0 : \rho_c = 0$ versus $H_1 : \rho_c > 0$. Similarly, both permutation tests used 5000 resamples. If we are testing at level $\alpha = 0.05$, then only studentized permutation test failed to reject H_0 ($p = 0.11$). All the other three tests rejected H_0 and conclude there is a non-zero agreement between FGS and the expert in estimating FAGS. Based on the simulation results in Section 3, the result from studentized test is more reliable and we should conclude that there is no significant agreement between FGS and the expert.

5 | DISCUSSION

In this work, we present a robust concordance correlation permutation test for testing $H_0 : \rho_c = \rho_{c(0)}$. Conventional testing of the CCC relies on large sample approximations, which tends to have inflated Type I error rates when the sample size is small. This was illustrated in our simulations studies (Section 3). An alternative approach to hypothesis testing based on asymptotic approximations is to consider the corresponding permutation test. However, DiCiccio and Romano[21] have shown that the naive permutation test of Pearson's correlation coefficient does not control type I error under non-normality settings where two variables can be dependent but uncorrelated. Here we demonstrated that a naive permutation test for the CCC suffers a similar issue both theoretically and empirically. To solve this issue, we proposed a permutation test for the CCC based on appropriately studentized statistic following DiCiccio and Romano's approach,[21] which controls type I error even when sample size is as small as 10 and normality assumption is violated.

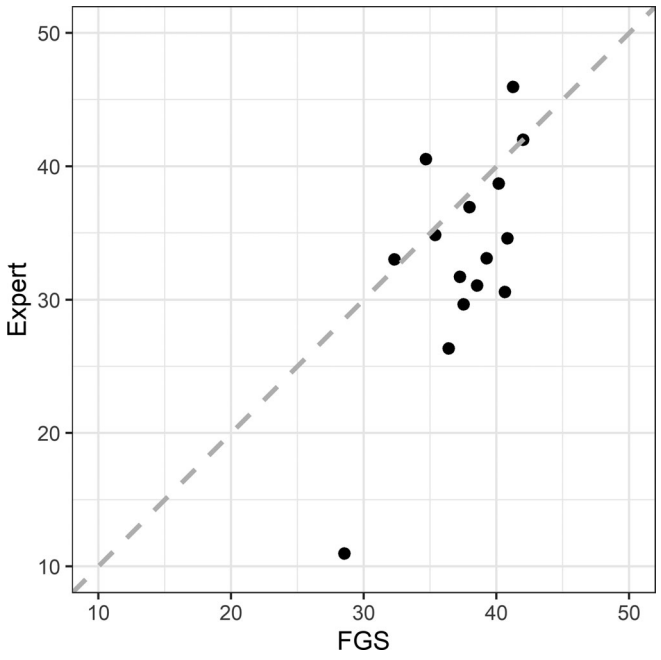

FIGURE 5 Scatter plot of echocardiographic imaging (EI) data with concordance line

TABLE 4 The p-values of testing $H_0 : \rho_c = 0$ versus $H_1 : \rho_c > 0$ for echocardiographic imaging (EI) data

Tests	*p* value
Asymptotic	0.0001
Fisher's Z	0.0003
Perm	0.0074
Stu Perm	0.1112

Importantly, while the original studentized permutation test can only handle tests of zero correlation coefficients,[24] we further extended the studentized permutation test for more general hypotheses. Similarly, the generalized testing procedure exhibited a robust type I error control under different scenarios. The proposed method may also enable the construction of confidence sets with better coverage probability by inverting the acceptance region, which still requires future investigation. Implementation of the method is available through the R package perk (https://github.com/hyu-ub/perk).

ACKNOWLEDGMENTS

This work was supported by Roswell Park Cancer Institute and National Cancer Institute (NCI) grant P30CA016056, NRG Oncology Statistical and Data Management Center grant U10CA180822 and IOTN Moonshot grant U24CA232979-01.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available.

ORCID

Han Yu https://orcid.org/0000-0001-6160-173X

REFERENCES

1. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology*. 2010;73(9):1167-1179.
2. Richard LJ, Koch Gary G. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
3. Jacob C. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213.
4. Martin BJ, Altman Douglas G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 327(8476):307-310.
5. Michael H. The design and analysis of clinical experiments. *J R Stat Soc: Ser A (Gen)*. 1987;150(4):400-400.
6. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45(1):255-268.
7. Andri S. DescTools: tools for descriptive statistics 2020. R package version 0.99.34.
8. Dai F. agRee: various methods for measuring agreement 2020. R package version 0.5-3.
9. Lluis CJ, Puig MJ. cccrm: concordance correlation coefficient for repeated (and non-repeated) measures 2015. R package version 1.2.1.
10. Barnhart Huiman X, Williamson JM. Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics*. 2001;57(3): 931-940.
11. Carrasco Josep L, Lluis J. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*. 2003;59(4):849-858.
12. Barnhart Huiman X, Michael H, Jingli S. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*. 2002;58(4):1020-1027.
13. Hutson AD. A multi-rater nonparametric test of agreement and corresponding agreement plot. *Comput Stat Data Anal*. 2010;54(1): 109-119.
14. Bertoli S, Posata A, Battezzati A, Spadafranca A, Testolin G, Bedogni G. Poor agreement between a portable armband and indirect calorimetry in the assessment of resting energy expenditure. *Clin Nutr*. 2008;27(2):307-310.
15. Guzmán-Venegas RA, Bralic MP, Cordero JJ, Cavada G, Araneda OF. Concordance of the location of the innervation zone of the tibialis anterior muscle using voluntary and imposed contractions by electrostimulation. *J Electromyogr Kinesiol*. 2016;27:18-23.
16. Anupa G, Jeevitha P, Bhat Muzaffer A, Bhagwan SJ, Jayasree S, Debabrata G. Endometrial stromal cell inflammatory phenotype during severe ovarian endometriosis as a cause of endometriosis-associated infertility: endometrial stromal cell behaviour during endometriosis. *Reprod BioMed Online*. 2020;41(4):623-639.
17. Krukowski Rebecca A, West Delia S, Marisha DC, et al. Are early first trimester weights valid proxies for preconception weight? *BMC Pregnancy Childbirth*. 2016;16(1):357.
18. Steichen Thomas J, Cox NJ. A note on the concordance correlation coefficient. *Stata J*. 2002;2(2):183-189.
19. Nicholas C, Thomas S. CONCORD: stata module for concordance correlation. 2007.
20. Williamson John M, Crawford Sara B, Hung-Mo L. Resampling dependent concordance correlation coefficients. *J Biopharm Stat*. 2007; 17(4):685-696.
21. DiCiccio CJ, Romano JP. Robust permutation tests for correlation and regression coefficients. *J Am Stat Assoc*. 2017;112(519):1211-1220.
22. Markus P, Edgar B, Frank K. Asymptotic permutation tests in general factorial designs. *J R Stat Soc: Ser B: Stat Methodol*. 2015;77: 461-473.
23. Sarah F, Frank K, Markus P. GFD: an R package for the analysis of general factorial designs. *J Stat Softw*. 2017;79(1):1-18.

24. Hutson AD. A robust Pearson correlation test for a general point null using a surrogate bootstrap distribution. *PLoS One*. 2019;14(5): e0216287.
25. Bowling LS, Sageman WS, O'Connor SM, Cole R, Amundson DE. Lack of agreement between measurement of ejection fraction by impedance cardiography versus radionuclide ventriculography. *Crit Care Med*. 1993;21(10):1523-1527.
26. Martin BJ, Altman Douglas G. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-160.
27. Geiser EA, Wilson DC, Wang DX, Conetta DA, Murphy JD, Hutson AD. Autonomous epicardial and endocardial boundary detection in echocardiographic short-axis images. *J Am Soc Echocardiogr*. 1998;11(4):338-348.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Hutson AD, Yu H. A robust permutation test for the concordance correlation coefficient. *Pharmaceutical Statistics*. 2021;20:696–709. https://doi.org/10.1002/pst.2101